HW1_DRL_Writeup_Bo-Yen Chang

Bo-Yen Chang | Cal ID Num: 3039662342 | CS 285 DRL

1.1 Inequality:

We need to show:

$$\sum_{st} |p_{\pi_\theta}(st) - p_{\pi^*}(st)| \leq 2T\varepsilon$$

By given:

$$\mathbb{E}_{p_{\pi^*}(s)}[\pi_\theta(a \neq \pi^*(s)|s)] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{p_{\pi^*}(st)}[\pi_\theta(at \neq \pi^*(st)|st)] \leq \varepsilon$$

Let's say, at the truth under :

$$\Delta_t = \pi_\theta(at \neq \pi^*(st)|st)$$

For each state (st), we have:

$$\mathbb{E}_{p_{\pi^*}(st)}[\Delta_t] \leq \varepsilon$$

Then, the difference in state distribution \(|p_{\pi_\theta}(st) - p_{\pi^*}(st)| \) for a particular state \(st \) can be bounded by the sum of differences in action choices between the two policies across time steps:

$$|p_{\pi_\theta}(st) - p_{\pi^*}(st)| \leq \sum_{t=1}^{T}\Delta_t$$

Now, using the union bound inequality:

$$\sum_{st} |p_{\pi_\theta}(st) - p_{\pi^*}(st)| \leq \sum_{t=1}^{T} \sum_{st} \Delta_t \leq T\varepsilon$$

Multiplying by 2, we get the results (we want):

$$\sum_{st} |p_{\pi_\theta}(st) - p_{\pi^*}(st)| \leq 2T\varepsilon$$

1.2 Expected Return:

(a)

If the reward only depends on the last state (Markov's), then:

$$r(st) = 0 \text{ for all } t < T \qquad\qquad J(\pi) = \mathbb{E}_{p_\pi(s_T)}[r(s_T)]$$

So, the difference is:

$$J(\pi^*) - J(\pi_\theta) = \mathbb{E}_{p_{\pi^*}(s_T)}[r(s_T)] - \mathbb{E}_{p_{\pi_\theta}(s_T)}[r(s_T)]$$

This can be bounded by:

$$|r_{\max}| \times |p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)|$$

from Part 1, we could know that it's complexity is

$$O(T\varepsilon)$$

(b)

For an arbitrary reward:

$$J(\pi) = \sum_{t=1}^{T} \mathbb{E}_{p_\pi(st)}[r(st)]$$

The difference is:

$$J(\pi^*) - J(\pi_\theta) = \sum_{t=1}^{T} \left(\mathbb{E}_{p_{\pi^*}(st)}[r(st)] - \mathbb{E}_{p_{\pi_\theta}(st)}[r(st)] \right)$$

For each term, the difference can be bounded by:

$$|r_{\max}| \times |p_{\pi^*}(st) - p_{\pi_\theta}(st)|$$

Given the result from part 1, summing up across all time steps:

$$J(\pi^*) - J(\pi_\theta) = O(T^2 \varepsilon)$$

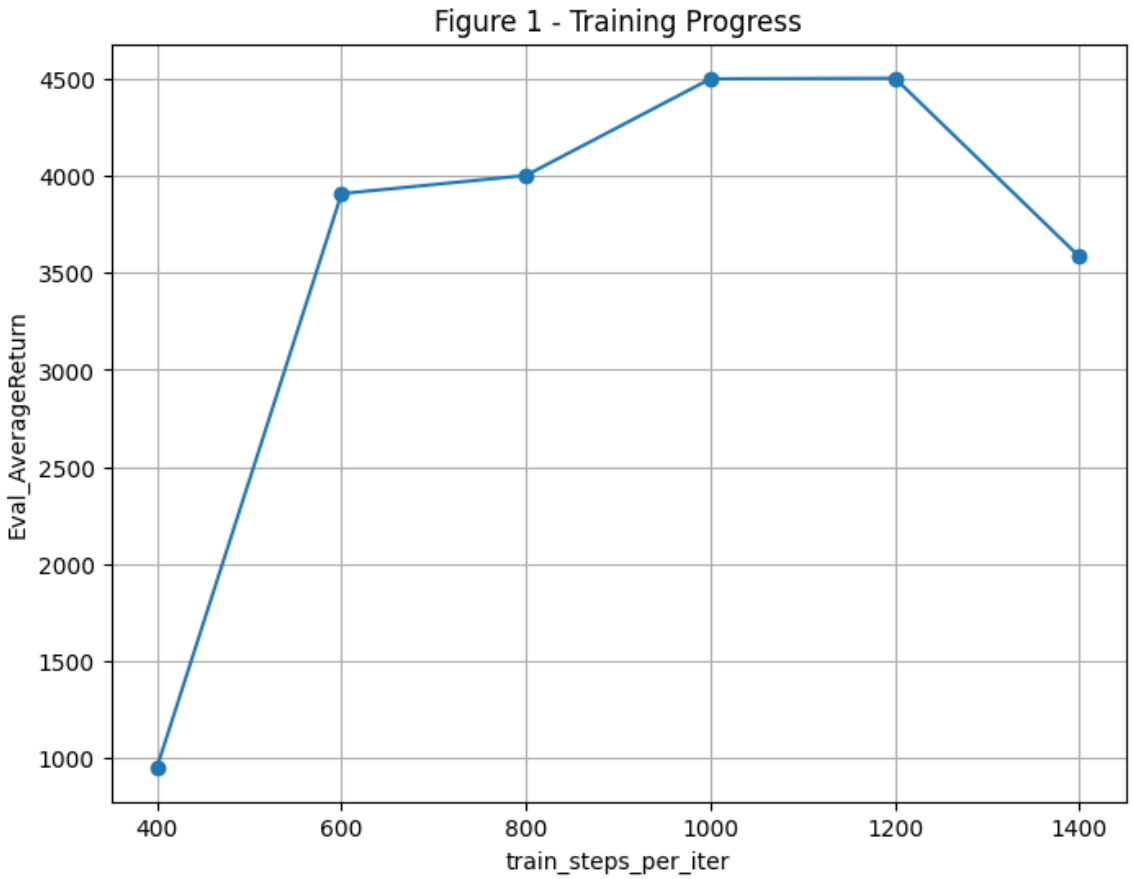
3.1

Table of Results

Tasks(env)	Eval_Average	Eval_Std	% of Expert
Ant	4500.7431640625	59.2415657043457	96.13%
Hopper	747.6747436523438	229.288269042968	20.11%

3.2

Figure 1



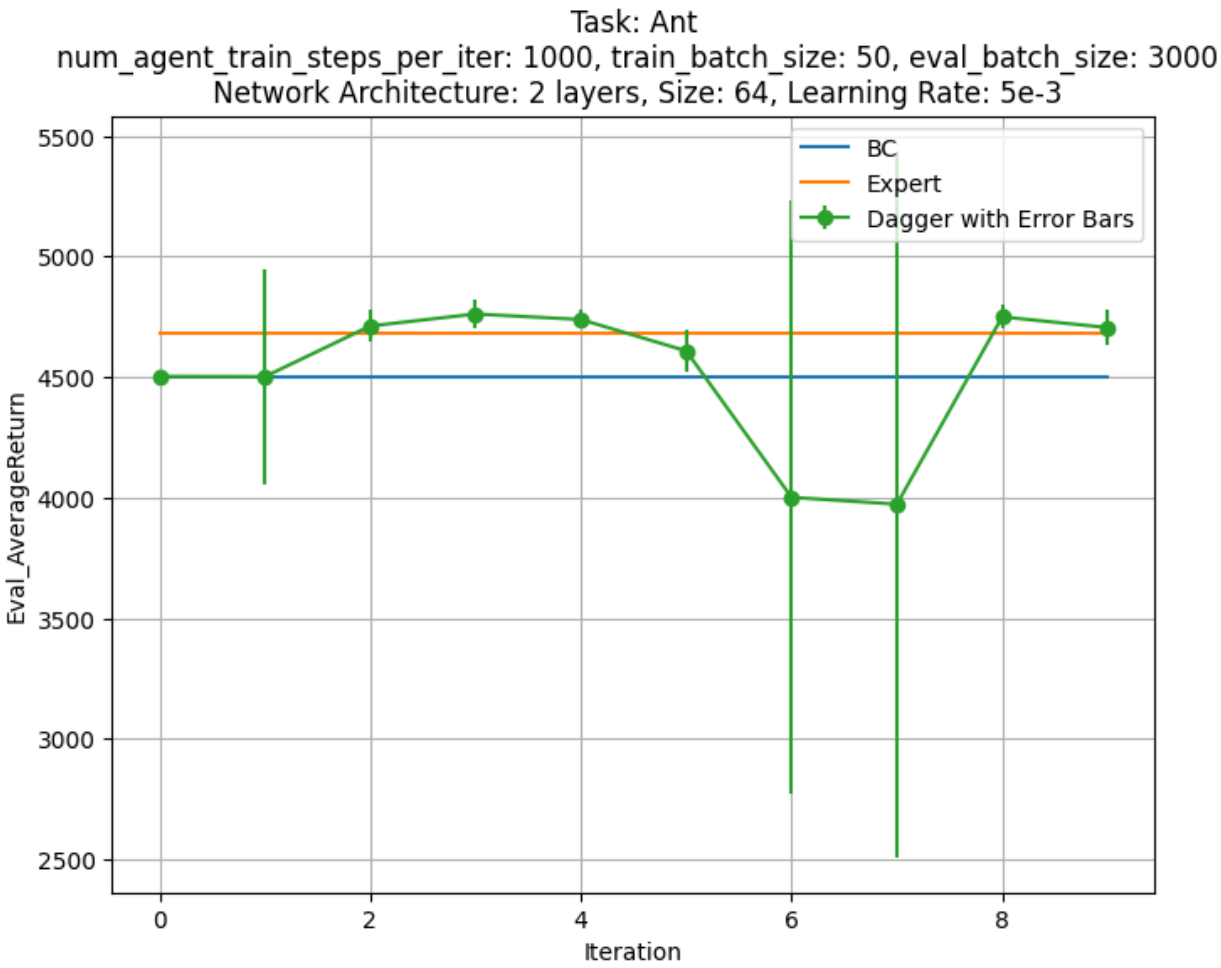
All of the tasks used a 2-layer network with each layer being size 64.

The training hyperparameters included:

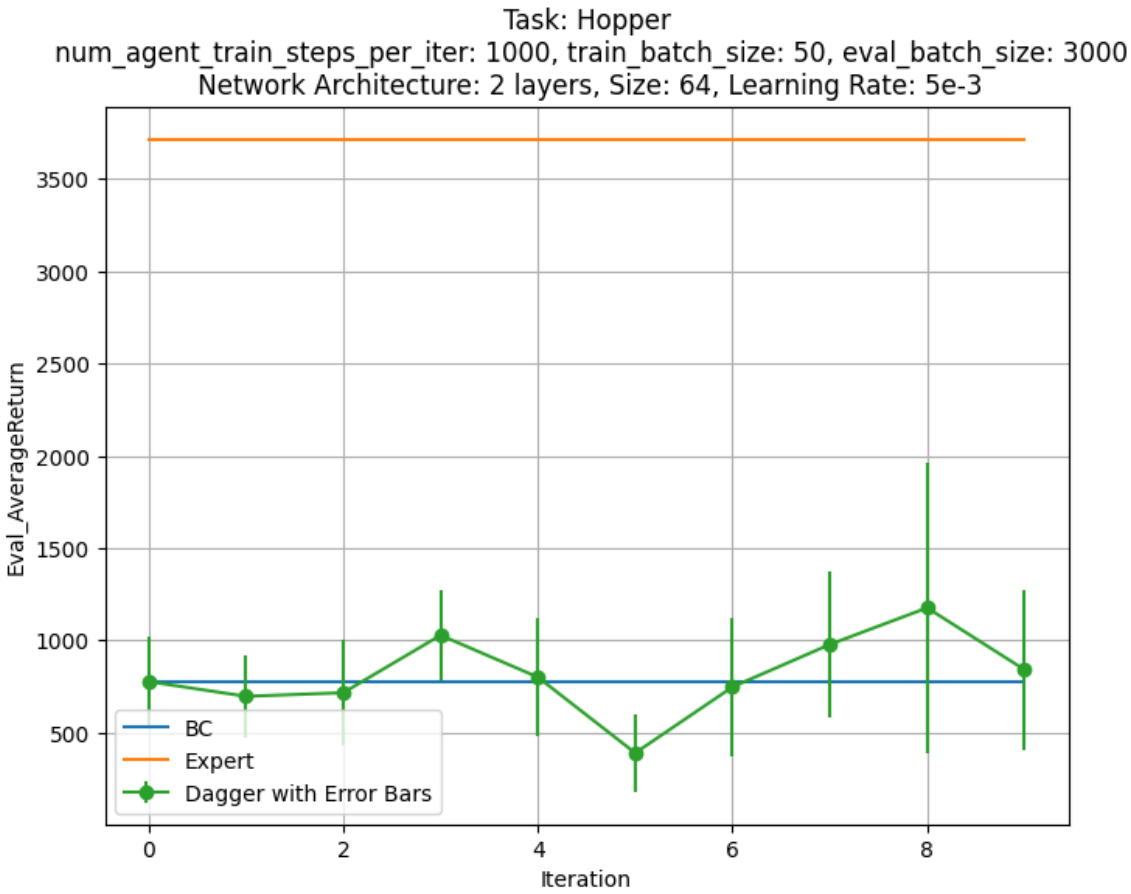
- train batch size (train_batch_size 50) = 50
- evaluation batch size (eval_batch_size) = 5000
- episode length (ep_len) = 1000
- number of training steps (num_agent_train_steps_per_iter)= 1000

4.1 & 4.2

Ant


Task: Ant
num_agent_train_steps_per_iter: 1000, train_batch_size: 50, eval_batch_size: 3000
Network Architecture: 2 layers, Size: 64, Learning Rate: 5e-3
BC
Expert
Dagger with Error Bars
Eval_AverageReturn
Iteration
5500
5000
4500
4000
3500
3000
2500
0
2
4
6
8

Hopper



All of the tasks used a 2-layer network with each layer being size (width) 64.

The training hyperparameters included:

- learning_rate: 5e-3
- train batch size (train_batch_size 50) = 50
- evaluation batch size (eval_batch_size) = 5000
- episode length (ep_len) = 1000
- number of training steps (num_agent_train_steps_per_iter)= 1000
- number of iterations: 10

5.1

❓ 1. How much time did you spend on each part of this assignment?

- **1.1** : 60 mins
- **1.2** : 60-90 mins
- **2** : 180-200 mins
- **3.1** : 100 mins (including debug)
- **3.2 :** 20 mins
- **4.1** : 35 mins
- **4.2** : 30 mins

5.2

❓ 2. Any additional feedback?

- I believe it would be beneficial to have a video where the GSI provides an overview of the code. This will allow students to ask questions upfront. Given the new code, it's challenging to grasp the structure in a short amount of time. Hence, an initial explanation, either in-person or through a pre-recorded video, would be immensely helpful.
- Additionally, the guidelines for submitting assignments seem a bit complicated. If there's a video demonstration showing the submission process, it can significantly enhance the efficiency of completing the assignments. This way, students can focus more on understanding the RL framework that the assignment aims to teach.